Importance sampling in signal processing applications

Rachel Ward

Abstract *Importance sampling* is a technique originating in Monte Carlo simulation whereby one samples from a different, *weighted* distribution, in order to reduce variance of the resulting estimator. More recently, variations of importance sampling have emerged as a means for reducing computational and sample complexity in different problems of modern signal processing. Here we review importance sampling as it is manifested in three such problems: stochastic optimization, compressive sensing, and low-rank matrix approximation. In keeping with a general trend in convex optimization towards the analysis of phase transitions for exact recovery, importance sampling in compressive sensing and low-rank matrix recovery can be used to effectively push the phase transition for exact recovery towards fewer measurements.

Key words: Complexity, Compressive sensing, Importance sampling, Matrix completion, Measurements, Stochastic gradient, Weighted sampling

Introduction

Importance sampling in simulation

The usual setup for importance sampling is in Monte Carlo simulation: one wants to compute an integral of the form $\int_{\mathscr{D}} f(x)p(x)dx$, where $p(x)$ is a probability density: $\int_{\mathscr{D}} p(x)dx = 1$. An easy and computationally efficient way to approximate such an integral is to consider the integral as an expectation, $\mu = \mathbb{E}(f(x)) = \int_{\mathscr{D}} f(x)p(x)dx$, and approximate the expectation as a sample average,

R. Ward (✉)
Department of Mathematics, University of Texas at Austin, Austin, TX, USA
e-mail: rward@math.utexas.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 4*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-20188-7_8

$$\int_{\mathscr{D}} f(x)dx \approx \frac{1}{m}\sum_{i=1}^{m} f(x_i), \quad x_i \sim p,$$

where the random variables x_i are independent and ideally distributed. Validity of this approximation is ensured by the law of large numbers, but the number of samples m needed for a given approximation accuracy grows with the variance of the random variable $f(x)$. In particular, if $f(x)$ is nearly zero on its domain $\mathscr{D}$ except in a region $A \subset \mathscr{D}$ for which $\mathbb{P}(x \in A)$ is small, then such standard Monte Carlo sampling may fail to have even one point inside the region A. It is clear intuitively that in this situation, we would benefit from getting some samples from the interesting or important region A. What *importance sampling* means is to sample from a different density $q(x)$ which overweights this region, rescaling the resulting quantity in order that the estimate remain unbiased.

More precisely, if x has probability density $p(x)$, then

$$\begin{aligned}\mu &= \mathbb{E}[f(x)] = \int_{\mathscr{D}} f(x)p(x)dx \\ &= \int_{\mathscr{D}} f(x)\frac{p(x)}{q(x)}q(x)dx = \mathbb{E}_q[f(x)w(x)],\end{aligned} \tag{1}$$

where $w(\cdot) \equiv \frac{p(\cdot)}{q(\cdot)}$ is the *weighting* function. By (1), the estimator

$$\widehat{\mu} = \frac{1}{m}\sum_{i=1}^{m} f(x_i)w(x_i), \qquad x_i \sim q, \tag{2}$$

is also an unbiased estimator for μ. The *importance sampling problem* then focuses on finding a biasing density $q(x)$ which overweights the important region close to an "optimal" way, at least such the *variance* of the importance sampling estimator is smaller than the variance of the general Monte Carlo estimate, so that fewer samples m are required to achieve a prescribed estimation error. In general, the density q^* with minimal variance σ^2_{q*} is proportional to $|f(x)|p(x)$, which is unknown a priori; still, there are many techniques for estimating or approximating this optimal distribution, see [31, Chapter 9].

Importance sampling beyond simulation

In recent times, probabilistic and stochastic algorithms have seen an explosion of growth as we move towards *bigger* data problems in *higher* dimensions. Indeed, we are often in the situation where at least one of the following is true:

1. Taking measurements is expensive, and we would like to reduce the number of measurements needed to reach a prescribed approximation accuracy

2. Optimizing over the given data is expensive, and we would like to reduce the number of computations needed to get within a prescribed tolerance of the optimal solution.

Importance sampling has proved to be helpful in both regimes. Whereas in simulation, importance sampling has traditionally been used for approximating *linear* estimates such as expectations/integrals, recent applications in signal processing and machine learning have considered importance sampling in approximating or even exactly recovering nonlinear estimates as well.

We consider here three case studies where the principle of importance sampling has been applied; this is by no means a complete list of all applications of importance sampling to machine learning and signal processing problems.

1. **Stochastic optimization:** Towards minimizing $F : \mathbb{R}^n \to \mathbb{R}$ of the form $F(x) = \sum_{i=1}^m f_i(x)$ via stochastic gradient descent, one iterates $x_{k+1} = x_k - \gamma w(i_k) \nabla f_{i_k}(x_k)$ with i_k randomly chosen from $\{1,2,\ldots,m\}$ so that

$$\mathbb{E}_{i_k}[x_{k+1}] = x_k - \gamma \sum_{i=1}^m \nabla f_i(x_k);$$

that is, one implements a full gradient descent update at each iteration, *in expectation*. Standard procedure is to sample indices from $\{1,2,\ldots,m\}$ uniformly, and the resulting convergence rate is limited by the *worst-case* Lipschitz constant associated with the component gradient functions. If however one has prior knowledge about the component Lipschitz constants, and has the liberty to draw indices proportionately to the associated Lipschitz constants, then the convergence rate of stochastic gradient can be improved so as to depend on the *average* Lipschitz constant among the components. This is in line with the principle of importance sampling: if ∇f_i has a larger Lipschitz constant, then this component is contributing more in content, and should be sampled with higher probability. We review some results of this kind in more detail below. For more details, see Section "Importance sampling in Stochastic Optimization".
2. **Compressive sensing:** Consider an orthonormal matrix $\Phi \in \mathbb{R}^{n\times n}$ (or $\Phi \in \mathbb{C}^{n\times n}$), along with a vector $x \in \mathbb{R}^n$. Then clearly

$$\Phi^* \Phi x = x;$$

moreover, if $\varphi_{i_k} \in \mathbb{R}^{1,n}$ is a randomly selected row from Φ, drawn such that row i is sampled with probability $p(i)$, then also

$$\mathbb{E}_p \left[\frac{1}{[p(i_k)]^2} \left(\varphi_{i_k}^* \varphi_{i_k} \right) \right] x = x.$$

Compressive sensing shows that if x is s-sparse, with $s \ll n$, then for certain orthonormal Φ, as few as $m \propto s \log^4(n)$ i.i.d. samples of the form $\langle \varphi_{i_k}, x \rangle$ can suffice to *exactly* recover x as the solution to a convex optimization program. For instance, such results hold if all of the rows of Φ are "equally important" (i.e., Φ

has uniformly bounded entries), and if rows are drawn i.i.d. uniformly from Φ. One may also incorporate importance sampling: if rows are drawn i.i.d. proportionately to their squared Euclidean norm, and if the *average* Euclidean row norm is small, then $m \propto s\log^4(n)$ i.i.d. samples still suffice for exact reconstruction. For more details, see Section "'Importance sampling in compressive sensing'.

3. **Low-rank matrix approximations:** Consider a matrix $M \in \mathbb{R}^{n_1 \times n_2}$ of rank $r \ll \min\{n_1, n_2\}$, and a subset $\Omega \subset [n_1] \times [n_2]$ of $|\Omega| = m$ revealed entries $M_{i,j}$. If the entries are revealed as i.i.d. draws where $\mathbf{Prob}[(i,j)] = p_{i,j}$, then $\mathbb{E}\left[\frac{1}{p_{i,j}} M_{i,j}\right] = M$. Importance sampling here corresponds to putting more weight $p_{i,j}$ on "important" entries in order to exactly recover M using fewer samples. We will see that if entries are drawn from a weighted distribution based on matrix *leverage scores*, then $m = r\log^2(\max\{n_1, n_2\})$ revealed entries suffices for M to be exactly recoverable as the solution to a convex optimization problem.

Importance sampling in Stochastic Optimization

Gradient descent is a standard method for solving unconstrained optimization problems of the form

$$\min_{x \in \mathbb{R}^n} F(x); \tag{3}$$

gradient descent proceeds as follows: initialize $x_0 \in \mathbb{R}^n$, and iterate along the direction of the negative gradient of F (the direction of "steepest descent") until convergence

$$x_{k+1} = x_k - \gamma_k \nabla F(x_k). \tag{4}$$

Here γ_k is the step-size which may change at every iteration. For optimization problems of very big size, however, even a full gradient computation of the form $\nabla F(x_k)$ can require substantial computational efforts and full gradient descent might not be feasible. This has motivated recent interest in random coordinate descent or stochastic gradient methods (see [3, 28, 29, 35, 36, 40], to name just a few), where one descends along gradient directions which are cheaper to compute. For example, suppose that F to be minimized is differentiable and admits a decomposition of the form

$$F(x) = \sum_{i=1}^{m} f_i(x). \tag{5}$$

Since $\nabla F(x) = \sum_{i=1}^{m} \nabla f_i(x)$, a full gradient computation involves computing all m gradients $\nabla f_i(x)$; still, one could hope to get *close* to the minimum, at a much smaller expense, by instead selecting a single index i_k at random from $\{1, 2, \ldots, m\}$ at each iteration. This is the principle behind *stochastic gradient* descent.

(5) **Stochastic Gradient (SG)**

Consider the minimization of $F : \mathbb{R}^n \to \mathbb{R}$ of the form $F(x) = \sum_{i=1}^m f_i(x)$.
Choose $x_0 \in \mathbb{R}^n$.
For $k \geq 1$ iterate until convergence criterion is met:

1. Choose $i_k \in [m]$ according to the rule $\mathbf{Prob}[i_k = k] = w(k)$
2. Update $x_{k+1} = x_k - \gamma \frac{1}{w(i_k)} \nabla f_{i_k}(x_k)$.

We have set the step-size γ to be constant for simplicity. Note that with the normalization in the update rule,

$$\begin{aligned} \mathbb{E}^{(w)}[x_{k+1}] &= x_k - \gamma \sum_{i=1}^m \nabla f_{i_k}(x_k) \\ &= x_k - \gamma \nabla F(x_k). \end{aligned} \tag{6}$$

Thus, we might hope for convergence *in expectation* of such stochastic iterations to the minimizer of (5) under similar conditions guaranteeing convergence of full gradient descent, namely, when F is convex (so that every minimizer is a global minimizer) and ∇F is Lipschitz continuous [30]. That is, we will assume

1. F is convex with convexity parameter $\mu = \mu(F) \geq 0$: for any x and y from $\mathbb{R}^n$ we have
$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{1}{2}\mu \|y - x\|^2. \tag{7}$$
When $\mu > 0$ strictly, we say that F is μ-strongly convex.
2. The component functions f_i are continuously differentiable and satisfy
$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|y - x\|, \qquad i = 1, 2, \ldots, m, \quad x, y \in \mathbb{R}^n. \tag{8}$$
We refer to L_i as the *Lipschitz constant* of ∇f_i.

The default sampling strategy in stochastic gradient methods is to sample uniformly, taking $w(i) = \frac{1}{m}$ in (5). In cases where the component functions f_i are only observed sequentially or in a streaming fashion, one does not have the freedom to choose a different sampling strategy. But if one *does* have such freedom, and has prior knowledge about the distribution of the Lipschitz constants L_i associated with the component function gradients, choosing probabilities $w(i) \propto L_i$ can significantly speed up the convergence rate of stochastic gradient. This is in line with the principle of importance sampling: if ∇f_i has a larger Lipschitz constant, it is contributing more in content, and should be sampled with higher probability. We review some results of this kind in more detail below.

Stochastic Gradient (SG) with Importance Sampling

For strongly convex functions, a central quantity in the analysis of stochastic descent is the *conditioning* of the problem, which is, roughly speaking, the ratio of the Lipschitz constant to the parameter of strong convexity. Recall that for a convex quadratic $F(x) = \frac{1}{2}x'Hx$, the Lipschitz constant of the gradient is given by the maximal eigenvalue of the Hessian H while the parameter of strong convexity is given by its minimal eigenvalue, and so in this case the conditioning reduces to the condition number of the Hessian matrix. In the general setting where $F(x) = \sum_{i=1}^m f_i(x)$ is strongly convex, the Hessian can vary with x, and the results will depend on the Lipschitz constants L_i of the ∇f_i and not only of the aggregate ∇F.

In short: with importance sampling, the convergence rate of stochastic descent is proportional to the *average conditioning* $\overline{L}/\mu = \frac{1}{m}\sum_{i=1}^m L_i/\mu$ of the problem; without importance sampling, the convergence rate must depend on the *uniform conditioning* $\sup_i L_i/\mu$. Thus, importance sampling has the highest potential impact if the Lipschitz constants are highly variable. This is made precise in the following theorem from [26], which in the case of uniform sampling, improves on a previous result of [2].

Theorem 1. *Let each f_i be convex where ∇f_i has Lipschitz constant L_i, with $L_i \leq \sup L$, and let $F(x) = \mathbb{E}f_i(x)$ be μ-strongly convex. Set $\sigma^2 = \mathbb{E}\|\nabla f_i(x_*)\|^2$, where $x_* = \text{argmin}_x F(x)$. Suppose that $\gamma \leq \frac{1}{\mu}$. Then the SG iterates in (5) satisfy:*

$$\mathbb{E}\|x_k - x_*\|^2 \leq \left[1 - 2\gamma\mu(1-\gamma\sup L)\right]^k \|x_0 - x_*\|^2 + \frac{\gamma\sigma^2}{\mu(1-\gamma\sup L)}. \tag{9}$$

where the expectation is with respect to the sampling of $\{i_k\}$ in (5).

The parameter σ^2 should be thought of as a 'residual' parameter measuring the extent to which the component functions f_i share a common minimizer. As a corollary of Theorem 1, if one pre-specifies a target accuracy $\varepsilon > 0$, then the optimal step-size $\gamma^* = \gamma^*(\varepsilon, \mu, \sigma^2, \sup L)$ is such that

$$k = 2\log(\varepsilon_0/\varepsilon)\left(\frac{\sup L}{\mu} + \frac{\sigma^2}{\mu^2\varepsilon}\right) \tag{10}$$

SG iterations suffice so that $\mathbb{E}\|x_k - x_*\|_2^2 \leq \varepsilon$,. See [26] for more details.

To see what this result implies for importance sampling, consider the stochastic gradient algorithm (5) with weights $w^{(k)}$. Then, when expectation is taken with respect to the sampling of $\{i_k\}$, we have $F(x) = \mathbb{E}f_i^{(w)}(x)$ where $f_i^{(w)} = \frac{1}{w^{(k)}}f_i$ has Lipschitz constant $L_i^{(w)} = \frac{1}{w(i)}L_i$. The supremum of $L_i^{(w)}$ is then given by:

$$\sup L_{(w)} = \sup_i L_i^{(w)} = \sup_i \frac{L_i}{w(i)}. \tag{11}$$

It is easy to verify that (11) is minimized by the weights

$$w(i) = \frac{L_i}{\overline{L}}, \quad \text{so that} \quad \sup L_{(w)} = \sup_i \frac{L_i}{L_i/\overline{L}} = \overline{L}. \tag{12}$$

Since μ is invariant to choice of weights, we find that in the "realizable" regime where $\sigma^2 = 0$, and hence $\sigma^2_{(w)} = 0$, then choosing the weights $w(i)$ as in (11) gives linear convergence with a linear dependence on the average conditioning $\overline{L}/\mu$, and a number of iterations,

$$k^{(w)} \propto \log(1/\varepsilon)\overline{L}/\mu,$$

to achieve a target accuracy ε. This strictly improves over the best possible results with uniform sampling, where the linear dependence is on the uniform conditioning $\sup L/\mu$ (see [26] for more details).

However, when $\sigma^2 > 0$, we get a potentially much *worse* scaling of the second term, by a factor of $\overline{L}/\inf L$:

$$\sigma^2_{(w)} = \mathbb{E}^{(w)}[\|\nabla f_i^{(w)}(x)\|_2^2] \leq \frac{\overline{L}}{\inf L}\sigma^2. \tag{13}$$

Fortunately, we can easily overcome this factor by sampling from a mixture of the uniform and fully weighted sampling, referred to as *partially biased sampling*. Using the weights

$$w(i) = \frac{1}{2}\frac{L_i}{\overline{L}} + \frac{1}{2}m,$$

we have

$$\sup L_{(w)} = \sup_i \frac{1}{\frac{1}{2} + \frac{1}{2}\cdot\frac{L_i}{\overline{L}}} L_i \leq 2\overline{L} \tag{14}$$

and

$$\sigma^2_{(w)} = \mathbb{E}\left[\frac{1}{\frac{1}{2} + \frac{1}{2}\cdot\frac{L_i}{\overline{L}}}\|\nabla f_i(x)\|_2^2\right] \leq 2\sigma^2. \tag{15}$$

In this sense, under the assumptions of Theorem 1, partially biased sampling will never be worse in terms of convergence rate than uniform sampling, up to a factor of 2, but can potentially have much better convergence.

Remark 1. An important example where all of these parameters have explicit forms is the *least squares problem*, where

$$F(x) = \frac{1}{2}\|Ax - b\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}(\langle a_i, x\rangle - b_i)^2, \tag{16}$$

with b an m-dimensional vector, A an $m \times n$ matrix with rows a_i, and $x_* = \arg\min_x \frac{1}{2}\|Ax - b\|_2^2$ is the least-squares solution. The Lipschitz constants of the components $f_i = \frac{m}{2}(\langle a_i, x\rangle - b_i)^2$ are $L_i = m\|a_i\|_2^2$, and the average Lipschitz constant is $\frac{1}{m}\sum_i L_i = \|A\|_F^2$ where $\|\cdot\|_F$ denotes the Frobenius norm. If A is full-rank and overdetermined, then F is strongly convex with strong convexity parameter $\mu = \|(A^TA)^{-1}\|_2^{-1}$, so that the average condition number is $\overline{L}/\mu = \|A\|_F^2\|(A^TA)^{-1}\|_2$.

Moreover, the residual is $\sigma^2 = m\sum_i \|a_i\|^2 |\langle a_i, x\rangle - b_i|^2$. Observe the bounds $\sigma^2 \leq n\|A\|_F^2 \sup_i |\langle a_i, x\rangle - b_i|^2$ and $\sigma^2 \leq m \sup_i \|a_i\|^2 \|Ax_* - b\|_2^2$.

Importance Sampling for SG in other regimes

Theorem 1 is stated for smooth and strongly convex objectives, and is particularly useful in the regime where the residual σ^2 is low, and the linear convergence term is dominant. But importance sampling can be incorporated into SG methods also in other regimes, and we now briefly survey some of these possibilities.

Smooth, Not Strongly Convex

When each component f_i is convex, non-negative, and has an L_i-Lipschitz gradient, but the objective $F(x)$ is not necessarily strongly convex, then after

$$k = O\left(\frac{(\sup L)\|x_*\|_2^2}{\varepsilon} \cdot \frac{F(x_*) + \varepsilon}{\varepsilon}\right) \tag{17}$$

iterations of SGD with an appropriately chosen step-size we will have $F(\bar{x}) \leq F(x_*) + \varepsilon$, where $\bar{x}$ is an appropriate averaging of the k iterates [43]. The relevant quantity here determining the iteration complexity is again $\sup L$. Furthermore, the dependence on the supremum is unavoidable and *cannot* be replaced with the average Lipschitz constant $\bar{L}$ [43]: if we sample gradients according to the uniform distribution, we must have a linear dependence on $\sup L$.

The only quantity in (17) that changes with a re-weighting is $\sup L$—all other quantities ($\|x_*\|_2^2$, $F(x_*)$, and the sub-optimality ε) are invariant to re-weightings. We can therefore replace the dependence on $\sup L$ with a dependence on $\sup L_{(w)}$ by using a weighted SGD as in (12). As we already calculated, the optimal weights are given by (12), and using them we have $\sup L_{(w)} = \bar{L}$. In this case, there is no need for partially biased sampling and we obtain that

$$k = O\left(\frac{\bar{L}\|x_*\|_2^2}{\varepsilon} \cdot \frac{F(x_*) + \varepsilon}{\varepsilon}\right) \tag{18}$$

iterations of weighed SGD updates (5) using the weights (12) suffice.

Non-Smooth Objectives

We now turn to non-smooth objectives, where the components f_i might not be smooth, but each component is G_i-Lipschitz. Roughly speaking, G_i is a bound on the first derivative (gradient) of f_i, while L_i is a bound on the second derivatives of f_i.

Here, the performance of SGD depends on the second moment $\overline{G^2} = \mathbb{E}[G_i^2]$. The precise iteration complexity depends on whether the objective is strongly convex or whether x_* is bounded, but in either case depends linearly on $\overline{G^2}$.

Using weighted SGD, we get linear dependence on:

$$\overline{G^2_{(w)}} = \mathbb{E}^{(w)}\left[(F_i^{(w)})^2\right] = \mathbb{E}^{(w)}\left[\frac{G_i^2}{w(i)^2}\right] = \mathbb{E}\left[\frac{G_i^2}{w(i)}\right], \tag{19}$$

where $F_i^{(w)} = G_i/w(i)$ is the Lipschitz constant of the scaled $f_i^{(w)}$. This is minimized by the weights $w(i) = G_i/\overline{G}$, where $\overline{G} = \mathbb{E}[G_i]$, yielding $\overline{G^2_{(w)}} = \overline{G}^2$. Using importance sampling, we reduce the dependence on $\overline{G^2}$ to a dependence on $\overline{G}^2$. It is helpful to recall that $\overline{G^2} = \overline{G}^2 + \mathrm{Var}[G_i]$. What we save is thus exactly the variance of the Lipschitz constants G_i. For more details, see [46].

Importance sampling in random coordinate descent

A related stochastic optimization problem is *randomized coordinate descent*, where one minimizes $F : \mathbb{R}^n \to \mathbb{R}$, not necessarily having the form $F(x) = \sum_{i=1}^m f_i(x)$, but still assumed to be strongly convex, by decomposing its gradient (5) into its *coordinate* directions

$$\nabla F(x) = \sum_{i=1}^{n} \nabla_i F(x)$$

and performing the stochastic updates:

1. Choose coordinate $i \in [n]$ according to rule $\mathbf{Prob}[i_k = k] = w(k)$
2. Update $x_{k+1} = x_k - \gamma \frac{1}{w(i_k)} \nabla_{i_k} F(x_k)$.

The motivation is that a coordinate directional derivative can be much simpler than computation of either function value, or a directional derivative along an *arbitrary* direction.

Actually, Theorem 1 can also be applied to this setting; its proof from [26] uses only that

$$\nabla F(x) = \mathbb{E}[\nabla f_i(x)], \tag{20}$$

and the fact that for, given any $x, y \in \mathbb{R}^n$,

$$\|\nabla f_i(x) - \nabla f_i(y)\|_2^2 \le L_i \langle x - y, \nabla f_i(x) - \nabla f_i(y)\rangle. \tag{21}$$

which follows from the assumption that f_i is smooth with Lipschitz continuous gradient by the so-called *co-coercivity Lemma*, see [26, Lemma A.1]. Note that (20) still

holds in the setting of randomized coordinate descent, and (21) holds if $F : \mathbb{R}^n \to \mathbb{R}$ has component-wise Lipschitz continuous gradient:

$$|\nabla_i F(x + he_i) - \nabla_i F(x)| \leq L_i |h|, \quad x \in \mathbb{R}^n,\ h \in \mathbb{R},\ i \in [n], \tag{22}$$

Under these assumptions, one may consider importance sampling for random coordinate descent with weights $w(k) = L_k / \sum_j L_j$, then we may apply Theorem 1 to get a linear convergence rate depending on $\overline{L}/\mu$ as opposed to $\sup L/\mu$. This is because coordinate descent falls into the *realizable* regime, as $\nabla_i F(x_*) = 0$ for each i, and hence also $\sigma^2 = \mathbb{E}\|(\nabla F)_i(x_*)\|^2 = 0$. Coordinate descent with importance sampling was considered before SG with importance sampling, originating in the works of [29] and [35]. One may consider the extension of randomized coordinate descent (8) to randomized *block* coordinate descent, descending in *blocks* of coordinates at a time. Then, the important Lipschitz constants are those associated with the *partial gradients* of F as opposed to the component-wise gradients [29].

Notes and extensions

Several aspects of importance sampling in stochastic optimization were not covered here, but we point out further results and references.

1. If the Lipschitz constants are not known a priori, then one could still consider doing importance sampling via *rejection sampling*, simulating sampling from the weighted distribution; this can be done by accepting samples with probability proportional to $L_i / \sup_j L_j$. The overall probability of accepting a sample is then $\overline{L}/\sup L_i$, introducing an additional factor of $\sup L_i/\overline{L}$, and thus again obtaining a linear dependence on $\sup L_i$. Thus, if we are only presented with samples from the uniform distribution, and the cost of obtaining the sample dominates the cost of taking the gradient step, we do not gain (but do not lose much either) from rejection sampling. We might still gain from rejection sampling if the cost of operating on a sample (calculating the actual gradient and taking a step according to it) dominates the cost of obtaining it and (a bound on) the Lipschitz constant.
2. All of the convergence results we stated in this section were with respect to the expected value. Nevertheless, all these rates extend to high probability results using Chebyshev's inequality. See [29] for more details.
3. Recently, several *hybrid* full-gradient/stochastic gradient methods have emerged which, as opposed to pure SG as in (5), have the advantage of progressively reducing the variance of the stochastic gradient with the iterations [19, 37, 41, 42], thus allowing convergence to the true minimizer. These algorithms can further be applied to the more general class of composite problems,

$$\text{minimize}_{x \in \mathbb{R}^n} \{P(x) = F(x) + R(x)\}, \tag{23}$$

where $F(x)$ is the average of many smooth component functions $f_i(x)$ whose gradients have Lipschitz constants L_i as in (5) and $R(x)$ is relatively simple but can

be non-differentiable. These algorithms have the added complexity of requiring a single pass over the data, all having complexity $O((n+\sup L/\mu)\log(1/\varepsilon))$. As shown in [45], importance sampling can also be applied in this more general setting to speed up convergence: sampling component functions proportional to their Lipschitz constants, this complexity bound becomes $O((n+\overline{L}/\mu)\log(1/\varepsilon))$.

4. An observation that is important not only for this chapter but also for the entire discussion on importance sampling is the computational cost of implementing a random counter, that is, given values $L_1, L_2, \ldots, L_m$, generate efficiently random integer numbers $i \in \{1,2,\ldots,m\}$ with probabilities

$$\mathbf{Prob}[i=k] = \frac{L_k}{\sum_{j=1}^{m} L_j}, \quad k = 1,2,\ldots,m. \tag{24}$$

Using a tree search algorithm [29], such a counter can be implemented with $\log(m)$ operations, and by generating one random number.

Importance sampling in compressive sensing

Introduction

The emerging area of mathematical signal processing known as *compressive sensing* is based on the observation that a signal which allows for an approximately sparse representation in a suitable basis or dictionary can be recovered from relatively few linear measurements via convex optimization, provided these measurements are sufficiently *incoherent* with the basis in which the signal is sparse [8, 10, 38]. In this section we will see how importance sampling can be used to enhance the incoherence between measurements and signal basis, again, allowing for recovery from fewer linear measurements.

We illustrate the power of importance sampling through two examples: compressed sensing imaging and polynomial interpolation. In compressed sensing imaging, coherence-based sampling provides a theoretical justification for empirical studies [23, 24] pointing to variable-density sampling strategies for improved MRI compressive imaging. In polynomial interpolation, coherence-based sampling implies that sampling points drawn from the Chebyshev distribution are better suited for the recovery of polynomials and smooth functions than uniformly distributed sampling points, aligning with classical results on Lagrange interpolation [5].

Before continuing, let us fix some notation. A vector $x \in \mathbb{C}^N$ is called s-sparse if $\|x\|_0 = \#\{x_j : x_j \neq 0\} \leq s$, and the best s-term approximation of a vector $x \in \mathbb{C}^N$ is the s-sparse vector $x_s \in \mathbb{C}^N$ satisfying $x_s = \inf_{u:\|u\|_0 \leq s} \|x-u\|_p$. Clearly, $x_s = x$ if x is s-sparse. Informally, x is called compressible if $\|x - x_s\|$ decays quickly as s increases.

Incoherence in compressive sensing

Here we recall sparse recovery results for structured random sampling schemes corresponding to *bounded orthonormal systems*, of which the partial discrete Fourier transform is a special case. We refer the reader to [15] for an expository article including many references.

Definition 1 (Bounded orthonormal system (BOS)). Let $\mathscr{D}$ be a measurable subset of $\mathbb{R}^d$.

- A set of functions $\{\psi_j : \mathscr{D} \to \mathbb{C}, \quad j \in [N]\}$ is called an *orthonormal system* with respect to the probability measure ν if $\int_{\mathscr{D}} \bar{\psi}_j(u)\psi_k(u)d\nu(u) = \delta_{jk}$, where δ_{jk} denotes the Kronecker delta.
- Let μ be a probability measure on $\mathscr{D}$. A *random sample* of the orthonormal system $\{\psi_j\}$ is the random vector $(\psi_1(T), \ldots, \psi_N(T))$ that results from drawing a sampling point T from the measure μ.
- An orthonormal system is said to be *bounded* with bound K if $\sup_{j\in[N]} \|\psi_j\|_\infty \leq K$.

Suppose now that we have an orthonormal system $\{\psi_j\}_{j\in[N]}$ and m random sampling points $T_1, T_2, \ldots, T_m$ drawn independently from some probability measure μ. Here and throughout, we assume that the number of sampling points $m \ll N$. As shown in [15], if the system $\{\psi_j\}$ is *bounded*, and if the probability measure μ from which we sample points is the orthogonalization measure ν associated with the system, then the (underdetermined) structured random matrix $A : \mathbb{C}^N \to \mathbb{C}^m$ whose rows are the independent random samples will be well conditioned, satisfying the so-called *restricted isometry property* [11] with nearly order-optimal restricted isometry constants with high probability. Consequently, matrices associated with random samples of bounded orthonormal systems have nice sparse recovery properties.

Proposition 1 (Sparse recovery through BOS). *Consider the matrix $A \in \mathbb{C}^{m\times N}$ whose rows are independent random samples of an orthonormal system $\{\psi_j,\ j \in [N]\}$ with bound $\sup_{j\in[N]} \|\psi_j\|_\infty \leq K$, drawn from the orthogonalization measure ν associated with the system. If the number of random samples satisfies*

$$m \gtrsim K^2 s \log^3(s)\log(N), \tag{25}$$

for some $s \gtrsim \log(N)$, then the following holds with probability exceeding $1 - N^{-C\log^3(s)}$: For each $x \in \mathbb{C}^N$, given noisy measurements $y = Ax + \sqrt{m}\eta$ with $\|\eta\|_2 \leq \varepsilon$, the approximation

$$x^\# = \arg\min_{z\in\mathbb{C}^N} \|z\|_1 \text{ subject to } \|Az - y\|_2 \leq \sqrt{m}\varepsilon$$

satisfies the error guarantee $\|x - x^\#\|_2 \lesssim \frac{1}{\sqrt{s}}\|x - x_s\|_1 + \varepsilon$.

An important special case of such a matrix construction is the *subsampled discrete Fourier matrix*, constructed by sampling $m \ll N$ rows uniformly at random from

the unitary discrete Fourier matrix $\Psi \in \mathbb{C}^{N\times N}$ with entries $\psi_{j,k} = \frac{1}{\sqrt{N}} e^{i2\pi(j-1)(k-1)}$. Indeed, the system of complex exponentials $\psi_j(u) = e^{i2\pi(j-1)u}$, $j \in [N]$, is orthonormal with respect to the uniform measure over the discrete set $\mathscr{D} = \{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\}$, and is bounded with optimally small constant $K = 1$. In the discrete setting, we may speak of a more general procedure for forming matrix constructions adhering to the conditions of Proposition 1: given any two unitary matrices Φ and Ψ, the composite matrix $\Phi^*\Psi$ is also a unitary matrix, and this composite matrix will have uniformly bounded entries if the orthonormal bases (ϕ_j) and (ψ_k), corresponding to the rows of Φ and Ψ, respectively, are *mutually incoherent:*

$$\mu(\Phi, \Psi) := \sqrt{N} \sup_{1\le j,k\le N} |\langle \phi_j, \psi_k \rangle| \le K. \tag{26}$$

Indeed, if Φ and Ψ are mutually incoherent, then the rows of $B = \sqrt{N}\Psi^*\Phi$ constitute a bounded orthonormal system with respect to the uniform measure on $\mathscr{D} = \{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\}$. Proposition 1 then implies a sampling strategy for reconstructing signals $x \in \mathbb{C}^N$ with assumed sparse representation in the basis Ψ, that is $x = \Psi b$ and $b \approx b_s$ (the s-sparse vector corresponding to its best s-term approximation), from a few linear measurements: form a sensing matrix $A \in \mathbb{C}^{m\times N}$ by sampling rows i.i.d. uniformly from an incoherent basis Φ, collect measurements $y = Ax + \eta$, $\|\eta\|_2 \le \varepsilon$, and solve the ℓ_1 minimization program,

$$x^{\#} = \arg\min_{z\in\mathbb{C}^N} \|\Psi^* z\|_1 \text{ subject to } \|Az - y\|_2 \le \sqrt{m}\varepsilon.$$

This scenario is referred to as *incoherent sampling*.

Importance sampling via local coherences

Consider more generally the setting where we aim to compressively sense signals $x \in \mathbb{C}^N$ with assumed sparse representation in the orthonormal basis $\Psi \in \mathbb{C}^{N\times N}$, but our sensing matrix $A \in \mathbb{C}^{m\times N}$ can only consist of rows from some fixed orthonormal basis $\Phi \in \mathbb{C}^{N\times N}$ that is not necessarily incoherent with Ψ. In this setting, we ask: *Given a fixed sensing basis Ψ and sparsity basis Φ, how should we sample rows of Ψ in order to make the resulting system as incoherent as possible?* We will answer this question by introducing the concept of *local coherence* between two bases as described in [21, 32], whereby in the discrete setting the coherences of individual elements of the sensing basis are calculated and used to derive the sampling strategy.

The following result quantifies how regions of the sensing basis that are more coherent with the sparsity basis should be sampled with higher density: they should be given more "importance". The following is essentially a generalization of Theorem 2.1 in [32], but for completeness, we include a short self-contained proof.

Theorem 2 (Sparse recovery via local coherence sampling). *Consider a measurable set $\mathscr{D}$ and a system $\{\psi_j,\ j \in [N]\}$ that is orthonormal with respect to a measure ν on $\mathscr{D}$ which has square-integrable local coherence,*

$$\sup_{j\in[N]} |\psi_j(u)| \le \kappa(u), \qquad \int_{u\in\mathscr{D}} |\kappa(u)|^2 \nu(u) du = B. \tag{27}$$

We can define the probability measure $\mu(u) = \frac{1}{B}\kappa^2(u)\nu(u)$ *on* $\mathscr{D}$. *Draw* m *sampling points* $T_1, T_2, \dots, T_m$ *independently from the measure* μ, *and consider the matrix* $A \in \mathbb{C}^{m\times N}$ *whose rows are the random samples* $\psi_j(T_k), j \in [N]$. *Consider also the diagonal preconditioning matrix* $\mathscr{P} \in \mathbb{C}^{m\times m}$ *with entries* $p_{k,k} = 1/\mu(T_k)$. *If the number of sampling points*

$$m \gtrsim B^2 s \log^3(s) \log(N), \tag{28}$$

for some $s \gtrsim \log(N)$, *then the following holds with probability exceeding* $1 - N^{-C\log^3(s)}$.

For each $x \in \mathbb{C}^N$, *given noisy measurements* $y = Ax + \sqrt{m}\eta$ *with* $\|\mathscr{P}\eta\|_2 \le \sqrt{m}\varepsilon$, *the approximation*

$$x^\# = arg\min_{z\in\mathbb{C}^N} \|z\|_1 \text{ subject to } \|\mathscr{P}Az - \mathscr{P}y\|_2 \le \sqrt{m}\varepsilon$$

satisfies the error guarantee

$$\|x - x^\#\|_2 \lesssim \frac{1}{\sqrt{s}}\|x - x_s\|_1 + \varepsilon.$$

The proof is a simple change-of-measure argument following the lines of standard importance sampling principle:

Proof. Consider the functions $Q_j(u) = \frac{\sqrt{B}}{\kappa(u)}\psi_j(u)$. The system $\{Q_j\}$ is bounded with $\sup_{j\in[N]} \|Q_j\|_\infty \le \sqrt{B}$, and this system is orthonormal on $\mathscr{D}$ with respect to the sampling measure μ:

$$\begin{aligned}
&\int_{u\in\mathscr{D}} \bar{Q}_j(u) Q_k(u) \mu(u) du \\
&= \int_{u\in\mathscr{D}} \left(\frac{1}{\kappa(u)}\bar{\psi}_j(u)\right)\left(\frac{1}{\kappa(u)}\psi_k(u)\right)\left(\kappa^2(u)\nu(u)\right) du \\
&= \int_{u\in\mathscr{D}} \bar{\psi}_j(u)\psi_k(u)\nu(u) du = \delta_{jk}.
\end{aligned} \tag{29}$$

Thus we may apply Proposition 1 to the system $\{Q_j\}$, noting that the matrix of random samples of the system $\{Q_j\}$ may be written as $\mathscr{P}A$.

In the discrete setting where $\{\psi_j\}_{j\in[N]}$ and $\{\phi_k\}$ are rows of unitary matrices Ψ and Φ, and ν is the uniform measure over the set $\mathscr{D} = \{0, \frac{1}{N}, \dots, \frac{N-1}{N}\}$, the integral in condition (27) reduces to a sum,

$$\sup_{k\in[N]} \sqrt{N}|\langle\psi_j, \phi_k\rangle| \le \kappa_j, \quad \frac{1}{N}\sum_{j=1}^{N} \kappa_j^2 = B. \tag{30}$$

This motivates the introduction of the local coherence of an orthonormal basis $\{\phi_j\}_{j=1}^N$ of $\mathbb{C}^N$ with respect to the orthonormal basis $\{\psi_k\}_{k=1}^N$ of $\mathbb{C}^N$:

Definition 2. The local coherence of an orthonormal basis $\{\phi_j\}_{j=1}^N$ of $\mathbb{C}^N$ with respect to the orthonormal basis $\{\psi_k\}_{k=1}^N$ of $\mathbb{C}^N$ is the function $\mu^{loc} = (\mu_j) \in \mathbb{R}^N$ defined coordinate-wise by

$$\mu_j = \sup_{1 \le k \le N} \sqrt{N} |\langle \varphi_j, \psi_k \rangle|.$$

We have the following corollary of Theorem 2.

Corollary 1. *Consider a pair of orthonormal basis (Φ, Ψ) with local coherences bounded by $\mu_j \le \kappa_j$. Let $s \ge 1$, and suppose that*

$$m \gtrsim s \left(\frac{1}{N} \sum_{j=1}^N \kappa_j^2 \right) \log^4(N).$$

Select m (possibly not distinct) rows of Φ^* independent and identically distributed from the multinomial distribution on $\{1,2,\ldots,N\}$ with weights $c\kappa_j^2$ to form the sensing matrix $A : \mathbb{C}^N \to \mathbb{C}^m$. Consider also the diagonal preconditioning matrix $\mathscr{P} \in \mathbb{C}^{m \times m}$ with entries $p_{k,k} = \frac{1}{\sqrt{c}\kappa_j}$. Then the following holds with probability exceeding $1 - N^{-C\log^3(s)}$: For each $x \in \mathbb{C}^N$, given measurements $y = Ax + \eta$, with $\|\mathscr{P}\eta\|_2 \le \sqrt{m}\varepsilon$, the approximation

$$x^\# = arg \min_{u \in \mathbb{C}^N} \|\Psi^* u\|_1 \text{ subject to } \|y - \mathscr{P}Au\|_2 \le \sqrt{m}\varepsilon$$

satisfies the error guarantee $\|x - x^\#\|_2 \lesssim \frac{1}{\sqrt{s}} \|\Psi^* x - (\Psi^* x)_s\|_1 + \varepsilon$.

Remark 2. Note that the local coherence not only influences the embedding dimension m, it also influences the sampling measure. Hence a priori, one cannot guarantee the optimal embedding dimension if one only has suboptimal bounds for the local coherence. That is why the sampling measure in Theorem 2 is defined via the (known) upper bounds κ and $\|\kappa\|_2$ rather than the (usually unknown) exact values μ_{loc} and $\|\mu_{loc}\|_2$, showing that local coherence sampling is *robust with respect to the sampling measure:* suboptimal bounds still lead to meaningful bounds on the embedding dimension.

We now present two applications where local-coherence sampling enables a sampling scheme with sparse recovery guarantees.

Remark 3. The $\log(N)^4$ factor in the required number of measurements, m, can be reduced to a single $\log(N)$ factor if one asks not for *uniform* sparse recovery (of the form "with high probability, this holds for all x") but rather a with-high probability result holding only for a particular x (of the form "for this x, recovery holds with high probability"). See [18] for more details.

Variable-density sampling for compressive sensing MRI

In Magnetic Resonance Imaging, after proper discretization, the unknown image (x_{j_1,j_2}) is a two-dimensional array in $\mathbb{R}^{n\times n}$, and allowable sensing measurements are two-dimensional Fourier transform measurements [1]:

$$\phi_{k_1,k_2} = \frac{1}{n}\sum_{j_1,j_2} x_{j_1,j_2} e^{2\pi i(k_1 j_1 + k_2 j_2)/n}, \quad -n/2+1 \le k_1,k_2 \le n/2.$$

Natural sparsity domains for images, such as discrete spatial differences, are not incoherent to the Fourier basis.

A number of empirical studies, including the very first papers on compressed sensing MRI, observed that image reconstructions from compressive frequency measurements could be significantly improved by variable-density sampling.

Note that lower frequencies are more coherent with wavelets and step functions than higher frequencies. In [21], the local coherence between the two-dimensional Fourier basis and bivariate Haar wavelet basis was calculated:

Proposition 2. *The local coherence between frequency ϕ_{k_1,k_2} and the bivariate Haar wavelet basis $\Psi = (\psi_I)$ can be bounded by*

$$\mu(\phi_{k_1,k_2},\Psi) \lesssim \frac{\sqrt{N}}{(|k_1+1|^2+|k_2+1|^2)^{1/2}}.$$

Note that this local coherence is *almost square integrable independent of discretization size n^2*, as

$$\frac{1}{N}\sum_{j=1}^{N}\mu_j^2 \lesssim \log(n).$$

Applying Corollary 1 to compressive MRI imaging, we then have

Corollary 2. *Let $n \in \mathbb{N}$. Let Ψ be the bivariate Haar wavelet basis and let $\Phi = (\phi_{k_1,k_2})$ be the two-dimensional discrete Fourier transform. Let $s \ge 1$, and suppose that $m \gtrsim s\log^5(N)$. Select m (possibly not distinct) frequencies (ϕ_{k_1,k_2}) independent and identically distributed from the multinomial distribution on $\{1,2,\dots,N\}$ with weights proportional to the inverse squared Euclidean distance to the origin, $\frac{1}{(|k_1+1|^2+|k_2+1|^2)}$, and form the sensing matrix $A : \mathbb{C}^N \to \mathbb{C}^m$. Then the following holds with probability exceeding $1-N^{-C\log^3(s)}$: for each image $x \in \mathbb{C}^{n\times n}$, given measurements $y = Ax$, the approximation*

$$x^{\#} = arg\min_{u\in\mathbb{C}^{n\times n}} \|\Psi^* u\|_1 \text{ subject to } \|\mathscr{D}y - Au\|_2 \le \varepsilon$$

satisfies the error guarantee $\|x-x^{\#}\|_2 \lesssim \frac{1}{\sqrt{s}}\|\Psi^* x - (\Psi^* x)_s\|_1 + \varepsilon$.

[1] The unknown might also be higher-dimensional, and is often 3-dimensional, but the ideas are analogous and we focus on the 2D example for simplicity.

Remark 4. This result was generalized to multidimensional wavelet and Fourier bases (not just two dimensions as considered above), and to any Daubechies wavelet basis in [20].

Remark 5. One can prove similar guarantees as in (2) using *total variation minimization* reconstruction, see [21, 25].

Sparse orthogonal polynomial expansions

Here we consider the problem of recovering polynomials g from m sample values $g(x_1), g(x_2) \dots, g(x_m)$, with sampling points $x_\ell \in [-1,1]$ for $\ell = 1, \dots, m$. If the number of sampling points is less or equal to the degree of g, then in general such reconstruction is impossible due to dimension reasons. However, the situation becomes tractable if we make a sparsity assumption. In order to introduce a suitable notion of sparsity, we consider the orthonormal basis of Legendre polynomials.

Definition 3. The (orthonormal) Legendre polynomials $P_0, P_1, \dots, P_n, \dots$ are uniquely determined by the following conditions:

- $P_n(x)$ is a polynomial of precise degree n in which the coefficient of x^n is positive,
- the system $\{P_n\}_{n=0}^{\infty}$ is orthonormal with respect to the normalized Lesbegue measure on $[-1,1]$: $\frac{1}{2}\int_{-1}^{1} P_n(x)P_m(x)dx = \delta_{n,m}, \quad n,m = 0,1,2,\dots$

Since the interval $[-1,1]$ is symmetric, the Legendre polynomials satisfy $P_n(x) = (-1)^n P_n(-x)$. For more information see [44].

An arbitrary real-valued polynomial g of degree $N-1$ can be expanded in terms of Legendre polynomials,

$$g(x) = \sum_{j=0}^{N-1} c_j P_j(x), \quad x \in [-1,1]$$

with coefficient vector $c \in \mathbb{R}^N$. The vector is s-sparse if $\|c\|_0 \leq s$. Given a set of m sampling points $(x_1, x_2, \dots, x_m)$, the samples $y_k = g(x_k)$, $k = 1, \dots, m$, may be expressed concisely in terms of the coefficient vector according to

$$y = \boldsymbol{\Phi} c,$$

where $\phi_{k,j} = P_j(x_k)$. If the sampling points $x_1, \dots, x_m$ are random variables, then the matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m \times N}$ is exactly the sampling matrix corresponding to random samples from the Legendre system $\{P_j\}_{j=1}^{N}$. This is not a bounded orthonormal system, however, as the Legendre polynomials grow like

$$|P_n(x)| \leq (n+1/2)^{1/2}, \quad -1 \leq x \leq 1.$$

Nevertheless the Legendre system does have bounded local coherence. A classic result from [44] follows.

Proposition 3. *For all $n > 0$ and for all $x \in [-1,1]$, $|P_n(x)| < \kappa(x) = 2\pi^{-1/2}(1-x^2)^{-1/4}$. Here, the constant $2\,\pi^{-1/2}$ cannot be replaced by a smaller one.*

Indeed, $\kappa(x)$ is a square integrable function proportional to the Chebyshev measure $\pi^{-1}(1-x^2)^{-1/2}$. We arrive at the following result for Legendre polynomial interpolation as a corollary of Theorem 2.

Corollary 3. *Let $x_1,\dots,x_m$ be chosen independently at random on $[-1,1]$ according to the Chebyshev measure $\pi^{-1}(1-x^2)^{-1/2}dx$. Let Ψ be the matrix with entries $\Psi_{k,j} = \sqrt{\pi/2}(1-x_k^2)^{1/4}P_n(x_k)$. Suppose that*

$$m \gtrsim s\log^3(N).$$

Consider the matrix $A \in \mathbb{C}^{m\times N}$ whose rows are independent random vectors $(\psi_j(X_k))$ drawn from the measure μ. If

$$m \gtrsim B^2 s\log^3(s)\log(N), \tag{31}$$

for some $s \gtrsim \log(N)$, then the following holds with probability exceeding $1 - N^{-C\log^3(s)}$. Let $\mathscr{D} \in \mathbb{C}^{m\times m}$ be the diagonal matrix with entries $d_{k,k} = \frac{1}{\mu(X_k)}$. For each $x \in \mathbb{C}^N$, given noisy measurements $y = Ax + \sqrt{m}\eta$ with $\|\mathscr{D}\eta\|_2 \le \sqrt{m}\varepsilon$, the approximation

$$x^{\#} = \arg\min_{u\in\mathbb{C}^N} \|u\|_1 \text{ subject to } \|\mathscr{D}Au - \mathscr{D}y\|_2 \le \sqrt{m}\varepsilon$$

satisfies the error guarantee $\|x-x^{\#}\|_2 \lesssim \frac{1}{\sqrt{s}}\|x-x_s\|_1 + \varepsilon$ where x_s is the best s-term approximation to x.

In fact, more general theorems exist: the Chebyshev measure is a universal sampling strategy for interpolation with any set of orthogonal polynomials [32]. An extension to the setting of interpolation with spherical harmonics, and more generally, to the eigenfunctions corresponding to smooth compact manifolds, can be found in [6, 32], respectively. For extensive numerical illustrations comparing Chebyshev vs. uniform sampling, also for high-dimensional tensor-product polynomial expansions, we refer the reader to [18].

Structured sparse recovery

Often, the prior of sparsity can be refined, and additional *structure* of the support set is known. In the MRI example where one senses with Fourier measurements signals which are sparse in Wavelets, the sparsity level will be higher for higher-order wavelets. One may consider sampling strategies based on a more refined notion of local coherence – based not only on $\mu_j = \sup_{1\le k\le N}\sqrt{N}|\langle\phi_j,\psi_k\rangle|$, but also coherences of sub-blocks $\mu_{j,B_k} = \sup_{k\in B_k}\sqrt{N}|\langle\phi_j,\psi_k\rangle|$. For more information, we refer the reader to the survey article [1] and the references therein.

In fact, we also have more information about the sparsity structure in the setting of function interpolation. It is well known that the smoothness of a function is reflected in the rate of decay of its Fourier coefficients / orthonormal Legendre polynomial coefficients, and vice versa. Thus, smooth functions have directional sparsity in their orthonormal polynomial expansions: low-order and low-degree polynomials are more likely to contribute to the representation. Another way to account for directional sparsity is in the reconstruction method itself. A more general theory of sparse recovery involves *weighted* ℓ_1 minimization as a reconstruction strategy, which serves as a *weighted* sparse prior, and the incorporation of importance sampling there, can be found in [33].

One of the motivating applications of sparse orthogonal polynomial expansions is toward the setting of *Polynomial Chaos expansions* in the area of *Uncertainty Quantification* (UQ), which involves high-dimensional expensive random inputs and modeling the output as having approximately sparse expansion in a tensorized orthogonal polynomial expansion. As shown in [18], in high dimensions, local coherence sampling strategy will depend on how high is the *dimension* compared to the maximal *order* of orthogonal polynomial considered; for higher-order models, Chebyshev sampling is a good strategy; for low-order, high-dimensional problems, uniform sampling outperforms Chebyshev sampling. For a detailed overview and more results, we refer the reader to [18].

Importance sampling in low-rank matrix recovery

Low-rank matrix completion

The task of *low-rank matrix completion* concerns the recovery of a low-rank matrix from a subset of its revealed entries, and nuclear norm minimization has emerged as an effective surrogate for this combinatorial problem. In fact, nuclear norm minimization can recover an arbitrary $n \times n$ matrix of rank r from $\mathcal{O}(nr\log^2(n))$ revealed entries, provided that revealed entries are drawn proportionally to the local row and column coherences (closely related to leverage scores) of the underlying matrix. Matrix completion has been the subject of much recent study due to its application in myriad tasks: collaborative filtering, dimensionality reduction, clustering, non-negative matrix factorization and localization in sensor networks. Clearly, the problem is ill-posed in general; correspondingly, analytical work on the subject has focused on the joint development of algorithms, and sufficient conditions under which such algorithms are able to recover the matrix.

If the true matrix is M with entries M_{ij}, and the set of observed elements is Ω, this method guesses as the completion the optimum of the convex program:

$$\begin{aligned} \min_{X} \quad & \|X\|_* \\ \text{s.t.} \quad & X_{ij} = M_{ij} \text{ for } (i,j) \in \Omega. \end{aligned} \tag{32}$$

where the "nuclear norm" $\|\cdot\|_*$ of a matrix is the sum of its singular values[2]. Throughout, we use the standard notation $f(n) = \Theta(g(n))$ to mean that $cg(n) \leq f(n) \leq Cg(n)$ for some positive constants c, C.

We focus on the setting where matrix entries are revealed from an underlying probability distribution. To introduce the distribution of interest, we first need a definition.

Definition 4. For an $n_1 \times n_2$ real-valued matrix M of rank r with SVD given by $U\Sigma V^\top$, the **local coherences**[3] – μ_i for any row i, and ν_j for any column j - are defined by the following relations

$$\begin{aligned} \left\|U^\top e_i\right\| &= \sqrt{\frac{\mu_i r}{n_1}} \quad , \quad i = 1, \ldots, n_1 \\ \left\|V^\top e_j\right\| &= \sqrt{\frac{\nu_j r}{n_2}} \quad , \quad j = 1, \ldots, n_2. \end{aligned} \tag{33}$$

Note that the μ_i, ν_js are non-negative, and since U and V have orthonormal columns we always have $\sum_i \mu_i r/n_1 = \sum_j \nu_j r/n_2 = r$.

The following theorem is from [13].

Theorem 3. *Let $M = (M_{ij})$ be an $n_1 \times n_2$ matrix with local coherence parameters $\{\mu_i, \nu_j\}$, and suppose that its entries M_{ij} are observed only over a subset of elements $\Omega \subset [n_1] \times [n_2]$. There are universal constants $c_0, c_1, c_2 > 0$ such that if each element (i, j) is independently observed with probability p_{ij}, and p_{ij} satisfies*

$$\begin{aligned} p_{ij} &\geq \min\left\{ c_0 \frac{(\mu_i + \nu_j) r \log^2(n_1 + n_2)}{\min\{n_1, n_2\}} \ , \ 1 \right\}, \\ p_{ij} &\geq \frac{1}{\min\{n_1, n_2\}^{10}}, \end{aligned} \tag{34}$$

then M is the unique optimal solution to the nuclear norm minimization problem (32) with probability at least $1 - c_1(n_1 + n_2)^{-c_2}$.

We will refer to the sampling strategy (34) as *local coherence* sampling. Note that the expected number of observed entries is $\sum_{i,j} p_{ij}$, and this satisfies

$$\begin{aligned} \sum_{i,j} p_{ij} &\geq \max\left\{ c_0 \frac{r \log^2(n_1 + n_2)}{\min\{n_1, n_2\}} \sum_{i,j} (\mu_i + \nu_j), \sum_{i,j} \frac{1}{n^{10}} \right\} \\ &= 2c_0 \max\{n_1, n_2\} r \log^2(n_1 + n_2), \end{aligned}$$

[2] This becomes the trace norm for positive-definite matrices. It is now well recognized to be a convex surrogate for rank minimization.

[3] In the matrix sparsification literature [4, 14] and beyond, the quantities $\left\|U^\top e_i\right\|^2$ and $\left\|V^\top e_j\right\|^2$ are referred to as the *leverage scores* of M.

independent of the coherence, or indeed any other property, of the matrix. Hoeffding's inequality implies that the actual number of observed entries sharply concentrates around its expectation, leading to the following corollary:

Corollary 4. *Let $M = (M_{ij})$ be an $n_1 \times n_2$ matrix with local coherence parameters $\{\mu_i, \nu_j\}$. Draw a subset of its entries by local coherence sampling according to the procedure described in Theorem 3. There are universal constants $c_1', c_2' > 0$ such that the following holds with probability at least $1 - c_1'(n_1+n_2)^{-c_2'}$: the number m of revealed entries is bounded by*

$$m \leq 3c_0 \max\{n_1, n_2\} r \log^2(n_1+n_2),$$

and M is the unique optimal solution to the nuclear norm minimization program (32).

(A) Roughly speaking, the condition given in (34) ensures that entries in important rows/columns (indicated by large local coherences μ_i and ν_j) of the matrix should be observed more often. Note that Theorem 3 only stipulates that an *inequality* relation hold between p_{ij} and $\{\mu_i, \nu_j\}$. This allows for there to be some discrepancy between the sampling distribution and the local coherences. It also has the natural interpretation that the more the sampling distribution $\{p_{ij}\}$ is "aligned" to the local coherence pattern of the matrix, the fewer observations are needed.

(B) Sampling based on local coherences provides close to the optimal number of sampled elements required for exact recovery (when sampled with any distribution). In particular, recall that the number of degrees of freedom of an $n \times n$ matrix with rank r is $2nr(1 - r/2n)$. Hence, regardless how the entries are sampled, a minimum of $\Theta(nr)$ entries is required to recover the matrix. Theorem 3 matches this lower bound, with an additional $O(\log^2(n))$ factor.

(C) Theorem 3 is from [13] and improves on the first results of matrix completion [7, 9, 17, 34] which assumed uniform sampling and incoherence i.e. every $\mu_i \leq \mu_0$ and every $\nu_j \leq \mu_0$ – and an additional *joint incoherence parameter* μ_{str} defined by $\|UV^\top\|_\infty = \sqrt{\frac{r\mu_{str}}{n_1 n_2}}$. The proof of Theorem 3 involves an analysis based on bounds involving the *weighted* $\ell_{\infty,2}$ matrix norm, defined as the maximum of the appropriately weighted row and column norms of the matrix. This differs from previous approaches that use ℓ_∞ or unweighted $\ell_{\infty,2}$ bounds [12, 17]. In some sense, using the weighted $\ell_{\infty,2}$-type bounds is natural for the analysis of low-rank matrices, because the rank is a property of the rows and columns of the matrix rather than its individual entries, and the weighted norm captures the relative importance of the rows/columns.

(D) If the column space of M is incoherent with $\max_i \mu_i \leq \mu_0$ and the row space is arbitrary, then one can randomly pick $\Theta(\mu_0 r \log n)$ rows of M and observe all their entries, and compute the local coherences of the space spanned by these rows. These parameters will be equal to the ν_j's of M with high probability. Based on these values, we can perform non-uniform sampling according to (34) and *exactly* recover M. Note that this procedure does not require any prior knowledge about the

local coherences of M. It uses a total of $\Theta(\mu_0 rn \log^2 n)$ samples. This was observed in [22].

Theorem 3 has some interesting consequences, discussed in detail in [13] and outlined below.

- Theorem 3 can be turned on its head, and used to quantify the benefit of *weighted* nuclear norm minimization over standard nuclear norm minimization, and provide a strategy for choosing the weights in such problems given non-uniformly distributed samples so as to reduce the sampling complexity of weighted nuclear norm minimization to that of standard nuclear norm minimization. In particular, these results can provide exact recovery guarantees for weighted nuclear norm minimization as introduced in [16, 27, 39], thus providing theoretical justification for its good empirical performance.
- Numerical evidence suggests that a two-phase adaptive sampling strategy, which assumes no prior knowledge about the local coherences of the underlying matrix M, can perform on par with the optimal sampling strategy in completing coherent matrices, and significantly outperform uniform sampling. Specifically, [13] considers a two-phase sampling strategy whereby given a fixed budget of m samples, one first draws a fixed proportion of samples uniformly at random, and then draw the remaining samples according to the local coherence structure of the resulting sampled matrix.

References

1. B. Adcock, A. Hansen, B. Roman, The quest for optimal sampling: computationally efficient, structure-exploiting measurements for compressed sensing. arXiv preprint (2014)
2. F. Bach, E. Moulines, Non-asymptotic analysis of stochastic approximation algorithms for machine learning, in *Advances in Neural Information Processing Systems* (2011)
3. L. Bottou, Large-scale machine learning with stochastic gradient descent, in *Proceedings of COMPSTAT'2010*, pp. 177–186, 2010
4. C. Boutsidis, M. Mahoney, P. Drineas, An improved approximation algorithm for the column subset selection problem, in *Proceedings of the Symposium on Discrete Algorithms*, pp. 968–977, 2009
5. L. Brutman, Lebesgue functions for polynomial interpolation: a survey. Ann. Numer. Math. **4**, 111– 127 (1997)
6. N. Burq, S. Dyatlov, R. Ward, M. Zworski, Weighted eigenfunction estimates with applications to compressed sensing. SIAM J. Math. Anal. **44**(5), 3481–3501 (2012)
7. E. Candès, B. Recht, Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717–772 (2009)
8. E.J. Candès, T. Tao, Near optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
9. E. Candès, T. Tao, The power of convex relaxation: near-optimal matrix completion. IEEE Trans. Inf. Theory **56**(5), 2053–2080 (2010)
10. E.J. Candès, T. Tao, J. Romberg, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
11. E.J. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)

12. Y. Chen, Incoherence-optimal matrix completion. arXiv preprint arXiv:1310.0154 (2013)
13. Y. Chen, S. Bhojanapalli, S. Sanghavi, R. Ward, Coherent matrix completion, in *Proceedings of the 31st International Conference on Machine Learning*, pp. 674–682, 2014
14. P. Drineas, M. Magdon-Ismail, M. Mahoney, D. Woodruff, Fast approximation of matrix coherence and statistical leverage. J. Mach. Learn. Res. **13**, 3475–3506 (2012)
15. S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Springer, Berlin, 2013)
16. R. Foygel, R. Salakhutdinov, O. Shamir, N. Srebro, Learning with the weighted trace-norm under arbitrary sampling distributions. arXiv:1106.4251 (2011)
17. D. Gross, Recovering low-rank matrices from few coefficients in any basis. IEEE Trans. Inf. Theory **57**(3), 1548–1566 (2011)
18. J. Hampton, A. Doostan, Compressive sampling of polynomial chaos expansions: convergence analysis and sampling strategies. J. Comput. Phys. **280**, 363–386 (2015)
19. R. Johnson, T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction. Adv. Neural Inf. Process. Syst. **26**, 315–323 (2013)
20. A. Jones, B. Adcock, A. Hansen, Analyzing the structure of multidimensional compressed sensing problems through coherence. arXiv preprint (2014)
21. F. Krahmer, R. Ward, Stable and robust sampling strategies for compressive imaging. IEEE Trans. Image Process. **23**(2), 612–622 (2014)
22. A. Krishnamurthy, A. Singh, Low-rank matrix and tensor completion via adaptive sampling. arXiv preprint. arXiv:1304.4672v2 (2013)
23. M. Lustig, D. Donoho, J. Pauly, Sparse mri: the application of compressed sensing for rapid mri imaging. Magn. Reson. Med. **58**(6), 1182–1195 (2007)
24. M. Lustig, D. Donoho, J. Santos, J. Pauly, Compressed sensing mri. IEEE Signal Process. Mag. **25**(2), 72–82 (2008)
25. D. Needell, R. Ward, Stable image reconstruction using total variation minimization. SIAM J. Imag. Sci. **6**(2), 1035–1058 (2013)
26. D. Needell, N. Srebro, R. Ward, Stochastic gradient descent and the randomized kaczmarz algorithm. arXiv preprint. arXiv:1310.5715 (2013)
27. S. Negahban, M. Wainwright, Restricted strong convexity and weighted matrix completion: optimal bounds with noise. J. Mach. Learn. Res. **13**, 1665–1697 (2012)
28. A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro, Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
29. Y. Nesterov, Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**(2), 341–362 (2012)
30. J. Nocedal, S.J. Wright, *Conjugate Gradient Methods* (Springer, Berlin, 2006)
31. A.B. Owen, *Monte Carlo Theory, Methods and Examples* (2013)
32. H. Rauhut, R. Ward, Sparse Legendre expansions via ℓ_1-minimization. J. Approx. Theory **164**, 517–533 (2012)
33. H. Rauhut, R. Ward, Interpolation via weighted ℓ_1 minimization. arXiv preprint. arXiv:1308.0759 (2013)
34. B. Recht, A simpler approach to matrix completion. arXiv preprint. arXiv:0910.0651 (2009)
35. P. Richtárik, M. Takáč, Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Math. Program. **144**(1), 1–38 (2014)
36. H. Robbins, S. Monrow, A stochastic approximation method. Ann. Math. Stat. **22**(22), 400–407 (1951)
37. N.L. Roux, M. Schmidt, F. Bach, A stochastic gradient method with an exponential convergence rate for finite training sets. Adv. Neural Inf. Process. Syst. **25**, 2672–2680 (2012)
38. M. Rudelson, R. Vershynin, On sparse reconstruction from Fourier and Gaussian measurements. Commun. Pure Appl. Math. **61**, 1025–1045 (2008)
39. R. Salakhutdinov, N. Srebro, Collaborative filtering in a non-uniform world: learning with the weighted trace norm. arXiv preprint. arXiv:1002.2780 (2010)
40. S. Shalev-Shwartz, N. Srebro, Svm optimization: inverse dependence on training set size, in *Proceedings of the 25th International Conference on Machine Learning*, pp. 928–935, 2008

41. S. Shalev-Shwartz, T. Zhang, Proximal stochastic dual coordinate ascent. arXiv:1211.2772 (2012)
42. S. Shalev-Shwartz, T. Zhang, Stochastic dual coordinate ascent methods for regularized loss minimization. J. Mach. Learn. Res. **14**, 567–599 (2013)
43. N. Srebro, K. Sridharan, A. Tewari, Smoothness, low noise and fast rates, in *Advances in Neural Information Processing Systems* (2010)
44. G. Szegö, *Orthogonal Polynomials* (American Mathematical Society, Providence, RI, 1939)
45. L. Xiao, T. Zhang, A proximal stochastic gradient method with progressive variance reduction. SIAM J. Optim. **24**(4), 2057–2075 (2014)
46. P. Zhao, T. Zhang, Stochastic optimization with importance sampling. arXiv:1401.2753 (2014)