Topic 8: Applications of Decoupling and Self-normalization

Victor H. de la Peña

Professor of Statistics, Columbia University

Artificial Intelligence Institute for Advances in Optimization
Georgia Institute of Technology 2024

Topics Preview

1. Application I: Boolean optimization

2. Application II: Linear Stochastic Bandits

In this section, we introduce the work by Makarychev and Sviridenko (2018, [3]), where the authors developed a tight decoupling inequality and used it to obtain the performance guarantee of approximation algorithms to minimize the cost of resources needed for a certain task.

Consider the following class of Boolean nonlinear programs:

$$\min_{\mathbf{x}\in\{0,1\}^n} \sum_j f_j\left(\sum d_{ij}x_i\right), \quad \text{s.t. } \mathbf{y} \in \mathcal{P} \tag{1}$$

where

- $\mathbf{x} = (x_1, \cdots, x_n) \in \{0,1\}^n$ consists of boolean variables to be optimized upon
- $\mathcal{P}$ is a polytope constraint of $\mathbf{x}$, i.e. a finite set of linear inequality constraints $\mathbf{a}_i^T\mathbf{x} \leq b_i$
- $d_{ij} \geq 0$ are known parameters of the system
- f_j is a non-decreasing convex cost function.

Let's start with a classical load balancing problem. Consider n jobs and m machines ($m < n$), and we need to decide which job each machine takes. Let x_{ij} denote whether job i is assigned to machine j, 1 being yes and 0 otherwise. Naturally we need each problem assigned only once, i.e.

$$\sum_{j=1}^{m} x_{ij} = 1, \forall i = 1, 2, ..., n.$$

If job i is assigned to machine j, the processing time is known to be $d_{ij} \geq 0$. The 'cost' of machine j is a polynomial of its total processing time: $(\sum_i d_{ij} x_{ij})^q$ for some $q > 1$ and we want to minimize the total cost. The problem can be formally written as

$$\min_{\mathbf{x}} \sum_{j=1}^{m} \left(\sum_{i=1}^{n} d_{ij} x_{ij} \right)^q, \quad \text{s.t.} \sum_{j=1}^{m} x_{ij} = 1, \forall i. \tag{2}$$

To match the form of (1), we can rewrite the inner sum to iterate on double subscript (i, k) and define $d_{i,j,k} = d_{ij}$ if $k = j$ and $d_{i,j,k} = 0$ otherwise. Problem (2) is then equivalent to

$$\min_{\mathbf{x}} \sum_{j=1}^{m} \left(\sum_{i,k} d_{ijk} x_{ik} \right)^q, \quad \text{s.t.} \ \sum_{j=1}^{m} x_{ij} = 1, \forall i. \tag{3}$$

Now we transform and relax problem (1). First notice that $\mathbf{y} \in \{0,1\}^n$ is uniquely determined by the index set S of 1s in it, so $\sum_i d_{ij} y_i = \sum_{i \in S} d_{ij}$. Our goal becomes finding S. Next, we introduce auxiliary indicator variables $\{z_{jS}\}_{j \le m, S \subset [n]}$ so that the original problem can be rewritten as

$$
\begin{aligned}
\min_{\substack{z_{jS} \\ j \in [m], S \subset [n]}} \quad & \sum_j \sum_{S \subset [n]} f_j \left(\sum_{i \in S} d_{ij} \right) z_{jS} \\
\text{s.t.} \quad & z_{jS} \in \{0,1\} \quad \forall j \in [m], S \subset [n] \\
& z_{1S} = \cdots = z_{mS} = z_S, \quad \forall S \subset [n] \\
& \sum_{S \subset [n]} z_S = 1 \\
& y_i = 1 \text{ iff } i \in S, \quad \forall i \\
& \mathbf{y} \in \mathcal{P}
\end{aligned}
\tag{4}
$$

This is a hard problem and solution is extremely slow. However, if we relax the constraint on z_{jS} then it becomes a linear programming problem:

$$\begin{aligned} \min_{\substack{z_{jS} \\ j\in[m], S\subset[n]}} \sum_{j} \sum_{S\subset[n]} f_j \left(\sum_{i\in S} d_{ij} \right) z_{jS} \\ \text{s.t. } z_{jS} \in [0,1] \quad \forall j \in [m], S \subset [n] \\ \sum_{S\subset[n]} z_{jS} = 1, \quad \forall j \\ \tilde{y}_i = \sum_{S:i\in S} z_{jS}, \quad \forall i,\ j \\ \tilde{\mathbf{y}} \in \mathcal{P} \end{aligned} \tag{5}$$

It can easily be verified that (5) is indeed a relaxation of (4) in that Recompile

1. Their target functions are the same,
2. All feasible solutions $\{z_{jS}, y_i\}$ of (5) are feasible to (4). More specifically, the chain $z_{1S} = z_{2S} = ... = z_{mS}$ is broken so that they can take different values, and boolean variables z_{jS} now can continuously take values from $[0, 1]$, so that y_i also become continuous variables $\tilde{y}_i$ in $[0, 1]$.

Note that (5) is indeed a linear programming problem as we declared, and it can be solved more quickly. Then one can randomly draw a boolean vector $\mathbf{y}$ based on $\tilde{\mathbf{y}}$, for example using independent $Bernoulli(\tilde{y}_i)$ distributions. However, one may ask: how well is such an approximation? The answer is in the following theorem (a modified version of Theorem 1.5 in Makarychev and Sviridenko 2015), which uses decoupling inequalities,

Theorem

Suppose $d_{ij} > 0$. *Let* $(\tilde{\mathbf{y}}^*, \mathbf{z}^*)$ *be a feasible solution of the relaxed problem* (5). *Assume each* Y_i *is drawn independently from Bernoulli*$(\tilde{y}_i)$. *Then we have*

$$\mathbb{E}\left[\sum_{j\in[k]} f_j\left(\sum_{i\in[n]} d_{ij} Y_i\right)\right] \leq \sum_{j\in[k]} A(f_j) \sum_{S\subseteq[n]} f_j\left(\sum_{i\in S} d_{ij}\right) z^*_{jS}, \qquad (6)$$

where $A(f) = \sup_{t>0} \mathbb{E}\left[f(tP)/f_j(t)\right]$, *with* P *a* $Poi(1)$ *variable.*
Particularly, since (5) *is a relaxation for* (1), *if* $(\tilde{\mathbf{y}}^*, \mathbf{z}^*)$ *is a* $(1+\epsilon)$*-approximately optimal solution to* (5), *then*

$$\mathbb{E}\left[\sum_{j\in[k]} f_j\left(\sum_{i\in[n]} d_{ij} Y_i\right)\right] \leq (1+\epsilon) \max_j \left(A(f_j)\right) IP \qquad (7)$$

where IP *is the optimal cost of the original problem* (1). *In other words,* $(Y_1, \cdots, Y_n)$ *is a* $(1+\epsilon)\max_j A(f_j)$*- sub-optimal solution of the original problem* (1)

In the load balancing problem, $f_j(x) = x^q$ for some $q > 1$. Then

$$A(f_j) = \sup_{t>0} \mathbb{E}\left[\frac{(tP)^q}{t^q}\right] = \mathbb{E}P^q$$

is the qth moment of $Poisson(1)$. When $q = 2$, $A(f_j) = 1$ so when $\tilde{\mathbf{y}}^*$ is a $(1+\epsilon)$-suboptimal solution of (5), $(Y_1, \cdots, Y_n)$ is a $(1+\epsilon)$-suboptimal solution of (1)!

The key to proving (6) is to bound $\mathbb{E} f_j(\sum_i d_{ij} Y_i)$, the expectation of a convex function of an independent sum. One can see that decoupling inequality plays an important role. In fact, the following decoupling inequality was developed by the authors and used to prove the above (6) (Theorem 5.3 in Makarychev and Sviridenko 2015), which is an extension to de la Peña [2].

Theorem

1. *Let $X_1, ..., X_n$ be nonnegative random variables, and let $Y_1, ..., Y_n$ be a complete decoupling of $X_1, ..., X_n$, i.e. Y_i are independent and have the same distribution as X_i. Let $P \sim Poisson(1)$, independent of X_i and Y_i's. Then for every convex function $\phi : \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{E}\left[\phi\left(\sum_{i=1}^{n} Y_i\right)\right] \leq \mathbb{E}\left[\phi\left(P\sum_{i=1}^{n} X_i\right)\right]. \tag{8}$$

2. *For every nonnegative function ϕ s.t. $\mathbb{E}[\phi(P)] < \infty$ and every positive ϵ, there exists n and random variables $X_1, ..., X_n$, $Y_1, ..., Y_n$ satisfying the assumptions in Part 1 such that*

$$\mathbb{E}\left[\phi\left(\sum_{i=1}^{n} Y_i\right)\right] \geq (1+\epsilon)\mathbb{E}\left[\phi\left(P\sum_{i=1}^{n} X_i\right)\right]. \tag{9}$$

Topics Preview

1. Application I: Boolean optimization

2. Application II: Linear Stochastic Bandits

In this section, we introduce another work by Abbasi-Yadkori, Pál and Szepesvári (2011) [1] in linear stochasic bandits. A confidence band was obtained by self normalization, and played a central role in the algorithm and risk analysis.

Here we introduce a paper by Abbasi et al (2011) [1]. The authors used pseudo-maximization to construct confidence sets, which in turn resulted in an algorithm with superior performance in terms of regret, compared to other algorithms at that time. Consider the following problem: assume at each round t, the player bets on d slot machines with money $X_t = (X_{t1}, ..., X_{td})$. The reward at time t is $Y_t = X_t^\top \theta^* + \eta_t$ where

- $\theta^* \in \mathbb{R}^d$ is a vector of unknown parameters,
- η_t is conditionally zero mean, i.e. $\mathbb{E}[\eta_t | \mathcal{F}_{t-1}] = 0$ where $\mathcal{F}_t = \sigma(\eta_1, \cdots, \eta_t)$

For simplicity we assume X_t is $\mathcal{F}_{t-1}$-measurable. The player's goal is to minimize $\sum_{i=1}^{n} X_t^\top \theta^*$ by properly choosing $\{X_t\}$ sequentially.
The authors consider an algorithm called Optimism in the Face of Uncertainty (OFUL) algorithm: it maintains a confidence set C_t of the unknown parameter $\theta*$. At round t, the learner

- Chooses $(X_t, \tilde{\theta}_t) := \arg \max_{x \in \mathbb{R}^d, \theta \in C_{t-1}} x^\top \theta$
- Receives new reward Y_t
- Updates the confidence set C_t.

Naturally, the construction of a confidence set is crucial to the algorithm, but it is difficult because the dependence structure of $\{X_t, Y_t\}_t$ is difficult to characterize. The following theorem provides a self-normalized concentration obtained from pseudo-maximization, and the concentration can be used to obtain a confidence ellipsoid C_t for θ^*.

Theorem (Abbasi-Yadkori et al. (2011))

Let $\{\eta_t\}_{t=1}^{\infty}$ *be a real-valued process, adapted to a filtration* $\{\mathcal{F}_t\}_{t=1}^{\infty}$*, and conditionally* R*-sub-Gaussian for some* $R \geq 0$*, i.e.*

$$\forall \lambda \in \mathbb{R}, \mathbf{E}[e^{\lambda \eta_t} | \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right)$$

let $\{X_t\}_{t=1}^{\infty}$ *be a* $\mathbb{R}^d$ *valued process such that* X_t *is* $\mathcal{F}_{t-1}$*-measurable. Assume* V *is a* $d \times d$ *positive definite matrix. For any* $t \geq 0$*, define*

$$\bar{V}_t = V + \sum_{s=1}^{t} X_s X_s^\top, \quad S_t = \sum_{s=1}^{t} \eta_s X_s$$

Then $\forall \delta > 0$*, with probability at least* $1 - \delta$*,* $\forall t \geq 0$*,*

$$S_t^\top \bar{V}_t^{-1} S_t \leq 2R^2 \log\left(\delta^{-1} \det(\bar{V}_t)^{1/2} \det(V)^{-1/2}\right)$$

Proof

The proof uses the previous lemma for matrix-normalized processes, combined with a stopping time argument. WLOG assume $R = 1$ (Othewise rescale η_t by η_t/R). Let τ be any stopping time. Denote $V_t := \sum_{s=1}^t X_s X_s^\top$ so that $\bar{V}_t = V + V_t$. We claim that (S_τ, V_τ) satisfy the canonical assumption, i.e.

$$\mathbb{E}e^{\lambda^\top S_\tau - \frac{1}{2}\lambda^\top V_\tau \lambda} \leq 1, \quad \forall \lambda \in \mathbb{R}^d$$

In fact, $M_t^\lambda := e^{\lambda^\top S_t - \frac{1}{2}\lambda^\top V_t \lambda}$ is a supermartingale (with additional definition $S_0 = V_0 = 0$ so $M_0^\lambda = 1$). To see this, notice that $M_t^\lambda = \prod_{s \leq t} e^{\lambda^\top X_s \eta_s - \frac{1}{2}\lambda^\top X_s X_s^\top \lambda}$ and the increment satisfies

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda^\top X_t \eta_t - \frac{1}{2}\lambda^\top X_t X_t^\top \lambda} \middle| \mathcal{F}_{t-1} \right] \\
&\leq e^{-\frac{1}{2}\lambda^\top X_t X_t^\top \lambda} \mathbb{E}\left[e^{\lambda^\top X_t \eta_t} \middle| \mathcal{F}_{t-1} \right] \\
&\leq e^{-\frac{1}{2}\lambda^\top X_t X_t^\top \lambda + \frac{1}{2}(\lambda^\top X_t)^2} \\
&= 1
\end{aligned}$$

Recalling Topic 6

Lemma (de la Peña, Klass & Lai, 2009 [4])

Let a random vector A and a symmetric, positive definite random matrix C satisfy the following canonical assumption

$$\mathbb{E} \exp\left(\theta^T A - \frac{1}{2}\theta^T C\theta\right) \leq 1, \quad \forall \theta \in \mathbb{R}^d. \tag{10}$$

Let V be a positive definite nonrandom matrix, then

$$\mathbb{E}\left[\sqrt{\frac{\det(V)}{\det(C+V)}} \exp\left(\frac{1}{2}A^T(C+V)^{-1}A\right)\right] \leq 1, \tag{11}$$

$$\mathbb{E} \exp\left(\frac{1}{4}A^T(C+V)^{-1}A\right) \leq \sqrt{\mathbb{E}\sqrt{\det(I+V^{-1}C)}}. \tag{12}$$

Proof Cont'd

This implies that M_t^λ is a supermartingale so M_τ^λ is a.s. well defined.
Clearly it follows that $\mathbb{E}M_\tau^\lambda \leq 1$.
By Equation (11), we have

$$\mathbb{E}\left[\sqrt{\frac{\det(V)}{\det(\bar{V}_\tau)}}e^{\frac{1}{2}S_\tau^\top \bar{V}_t^{-1}S_\tau}\right] \leq 1.$$

It follows that

$$\begin{aligned}
&\mathbb{P}\left(S_\tau^\top \bar{V}_\tau^{-1}S_\tau > 2\log\left(\delta^{-1}\det(\bar{V}_\tau)^{1/2}\det(V)^{-1/2}\right)\right) \\
&= \mathbb{P}\left(e^{\frac{1}{2}S_\tau^\top \bar{V}_\tau^{-1}S_\tau}\det(\bar{V}_\tau)^{-1/2}\det(V)^{1/2} > \delta^{-1}\right) \\
&\leq \delta\mathbb{E}\left[\sqrt{\frac{\det(V)}{\det(\bar{V}_\tau)}}e^{\frac{1}{2}S_\tau^\top \bar{V}_t^{-1}S_\tau}\right] \leq \delta
\end{aligned}$$

Proof Cont'd

Finally, to prove the union bound over t, let τ be the first time that the inequality fails, i.e.

$$\tau := \inf\{t \geq 0 : S_t^\top \bar{V}_t^{-1} S_t > 2\log\left(\delta^{-1}\det(\bar{V}_t)^{1/2}\det(V)^{-1/2}\right)\}$$

with the convention that $\inf \emptyset = +\infty$. Then

$$\begin{aligned}
&\mathbb{P}\left(\exists t \geq 0, S_t^\top \bar{V}_t^{-1} S_t > 2\log\left(\delta^{-1}\det(\bar{V}_t)^{1/2}\det(V)^{-1/2}\right)\right) \\
&= \mathbb{P}(\tau < \infty) \\
&= \mathbb{P}\left(S_\tau^\top \bar{V}_\tau^{-1} S_\tau > 2\log\left(\delta^{-1}\det(\bar{V}_\tau)^{1/2}\det(V)^{-1/2}\right), \tau < \infty\right) \\
&\leq \mathbb{P}\left(S_\tau^\top \bar{V}_\tau^{-1} S_\tau > 2\log\left(\delta^{-1}\det(\bar{V}_\tau)^{1/2}\det(V)^{-1/2}\right)\right) \\
&\leq \delta.
\end{aligned}$$

The theorem can be used to construct confidence sets in the following way: let $\hat{\theta}_t = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{Y}$ be the ridge estimator of θ^* using current information. $\forall x \in \mathbb{R}^d$, define $\|x\|_V := x^\top V x$ for positive definite V:

Theorem

Let $V = \lambda I$ in the definition of $\bar{V}_t$. Suppose $\|\theta^*\| \leq K$. Then $\forall \delta > 0$, with probability at least $1 - \delta$, $\forall t$, θ^* lies in the confidence set C_t

$$\left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{\bar{V}_t} \leq R \sqrt{2 \log \left(\delta^{-1} \det(\bar{V}_t)^{1/2} \det(\lambda \mathbf{I})^{-1/2} \right)} + \lambda^{1/2} K \right\}$$

The key is to connect $S_t = \sum \eta_s X_s$ and $\|\theta - \hat{\theta}_t\|_{\bar{V}_t}$.

References I

[1] Y. Abbasi-yadkori, D. Pál, and C. Szepesvári. "Improved Algorithms for Linear Stochastic Bandits". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor et al. Vol. 24. Curran Associates, Inc., 2011.

[2] V. H. de la Peña. "Bounds on the expectation of functions of martingales and sums of positive RVs in terms of norms of sums of independent random variables". In: *Proc. Amer. Math. Soc.* 108.1 (1990), pp. 233–239.

[3] K. Makarychev and M. Sviridenko. "Solving optimization problems with diseconomies of scale via decoupling". In: *J. ACM* 65.6 (2018), pp. 1–27.

References II

[4] Victor H de la Pena, Michael J Klass, and Tze Leung Lai. "Theory and applications of multivariate self-normalized processes". In: *Stochastic Processes and their Applications* 119.12 (2009), pp. 4210–4227.